# MILCA: Malaria Parasite Detection from Sample-Level Weak Labels

**Petru Manescu** P.MANESCU@UCL.AC.UK
**Delmiro Fernandez-Reyes** DELMIRO.FERNANDEZ-REYES@UCL.AC.UK
*Department of Computer Science, University College London*

**Editors:** Under Review for MIDL 2026

## Abstract

Malaria diagnosis requires the inspection of multiple image fields per sample. Training vision models for malaria parasite detection typically requires large numbers of expert-provided bounding boxes, which are costly to obtain and often impractical in real-world deployments. We introduce MILCA, a weakly supervised object detection framework that learns parasite localization from sample-level diagnostic labels, which are routinely recorded in clinical practice. MILCA combines Multiple Instance Learning (MIL) for sample classification with an iterative Class Activation (CA) Mapping procedure that yields coarse parasite pseudo-labels, which are further enriched with hard negatives from parasite-free samples. These pseudo-labels enable training a detector without any manual bounding-box supervision. Experiments on multiple microscopy datasets show that MILCA achieves reliable detection and counting performance under fully weak supervision, and that fine-tuning with only a small fraction of expert annotations provides substantial additional gains, outperforming supervised and pseudo-labeling baselines under the same or lower annotation budgets. By converting coarse, sample-level clinical labels into effective object-level supervision and leveraging a hybrid refinement scheme, MILCA provides a label-efficient route toward automated malaria parasite and detection and a general approach for weakly supervised blood film analysis.

**Keywords:** Malaria, Weak Supervision, Object Detection, Multiple Instance Learning, Class Activations

## 1. Introduction

Malaria remains a major global health challenge, disproportionately affecting children in low- and middle-income countries (LMICs) of the Global South (Organization et al., 2022). The current gold diagnosis standard is the visual inspection of Giemsa-stained Thick Blood Films (TBFs). Microscopists must identify extremely small, faint parasites among numerous artefacts such as platelets or staining debris, a demanding and time-consuming process that is difficult to scale in resource-limited regions (Organization and for Disease Control, 2010). Parasite density estimation further guides prognosis and treatment efficacy (Siahaan, 2018). Automated detection systems based on deep learning offer promise (Mehanian et al., 2017; Torres et al., 2018; Yang et al., 2019; Manescu et al., 2020; Chibuta and Acar, 2020; Kassim et al., 2021), yet nearly all rely on tens of thousands of manually annotated bounding boxes, and on segmentation-based region proposal techniques, an impractical requirement for widespread deployment.

By contrast, *sample-level* diagnostic labels (malaria-positive vs. malaria-negative) are routinely collected as part of standard clinical workflows and are inexpensive to acquire.

© 2026 CC-BY 4.0, P. Manescu & D.F.-R. .

TBF samples naturally form *bags* of image fields, suggesting a Multiple Instance Learning (MIL) formulation. MIL has been widely used in computational pathology to learn from slide-level labels (Campanella et al., 2019; Gadermayr and Tschuchnig, 2024) and has seen limited use in hematology (Sidhom et al., 2021; Manescu et al., 2023; Gao et al., 2023), but has not been explored for parasitic infection detection in TBFs, where object instances are tiny, numerous, and visually ambiguous.

MIL alone, however, produces only sample-level predictions and does not yield object-level localization, which is essential for parasite density estimation and clinical interpretability. Weakly supervised object detection (WSOD) methods based on Class Activation Maps (CAMs) have been used to derive localization cues from image-level labels (Zhang et al., 2021; Belharbi et al., 2025; Kniesel et al., 2025), but they have not been investigated for high-magnification bright-field TBF microscopy. Moreover, existing WSOD pipelines assume that each labeled image contains at least one target instance, whereas a malaria-positive sample may contain hundreds of fields and only a small fraction with parasites. For example, (Kniesel et al., 2025) assumes image-level labels for single electron-microscopy images, a modality not used in clinical workflows. In contrast, MILCA operates on clinically acquired brightfield thick blood films, where each sample consists of up to 100 fields and requires a MIL backbone rather than a single-image classifier to account for the sample-level diagnostic labels routinely produced in practice.

In addition to weakly supervised methods, semi-supervised learning (SemiSL) has been widely investigated to reduce annotation costs in medical imaging. SSL frameworks leverage both limited labeled data and large pools of unlabeled images, tipically through consistency regularization, pseudo-labeling, or teacher–student schemes (Chen et al., 2022; Bai et al., 2017; Madani et al., 2018; Xu et al., 2022; Shokrollahi et al., 2024). However, these methods require an initial supervised detector, a challenge in malaria microscopy, where only a handful of bounding boxes may be available. MILCA circumvents this limitation by constructing the initial detector entirely from weak sample-level supervision before incorporating limited strong labels.

Our main contributions are:

1. **MIL for malaria diagnosis.** We formulate thick blood film analysis as a Multiple Instance Learning problem operating directly on bags of high-magnification fields, enabling malaria positivity prediction from sample-level labels without candidate extraction, handcrafted preprocessing, or ROI selection.

2. **ICAM: a domain-adapted iterative CAM refinement.** Building on WSOD principles but adapted to small, dense malaria parasites in TBF, ICAM combines confidence-coupled erasure, decayed CAM accumulation, and hard-negative mining to obtain usable pseudo-annotations from MIL predictions.

3. A hybrid weak+strong supervision regime with strong generalization and high label efficiency. We show that detectors initialized purely from MILCA pseudo-annotations and fine-tuned with only a small fraction of manual labels outperform supervised and simple SemiSL baselines under the same—or even lower—annotation budgets. A multi-center evaluation further demonstrates robust cross-site generalization and improved label efficiency.

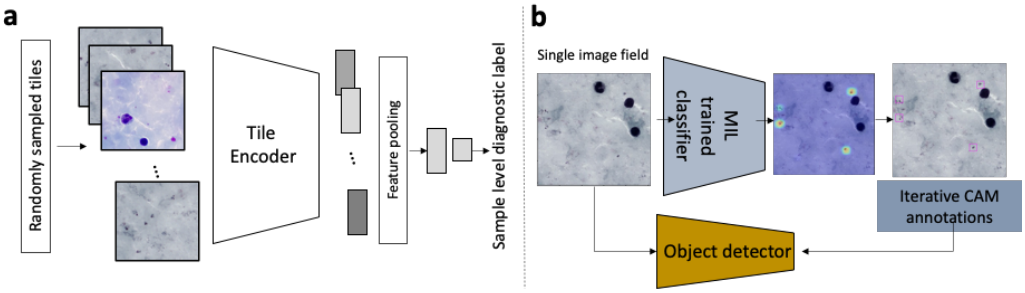



Figure 1: Brief overview of MILCA. (a) Multiple Instance Learning for TBF sample classification. (b) CAM-based generated annotations are used to train an object detector.

## 2. Methods

MILCA converts sample-level diagnostic labels into object-level supervision through three stages: (i) multiple instance learning (MIL) for sample classification, (ii) iterative CAM refinement (ICAM) for pseudo-annotation generation, and (iii) pseudo-label–driven detector training with optional fine-tuning (Fig. **??**).

### 2.1. Datasets

We trained and evaluated MILCA on three thick blood film (TBF) microscopy datasets collected from different geographic and clinical settings (Table 1). The Ibadan dataset (Manescu et al., 2020) contains 144 malaria-negative and 155 malaria-positive samples, each with approximately 100 high-resolution image fields acquired with a 100× oil-immersion objective. The Chittagong-1 dataset (Kassim et al., 2021) comprises 50 negative and 150 positive samples, with 20 image fields per sample at higher resolution. Finally, the Chittagong-2 dataset (Yang et al., 2019) includes 150 positive samples with over 80k expert-annotated parasites, and was used exclusively for the external evaluation of parasite detection performance. Together, these datasets span substantial variation in staining, imaging hardware, magnification, and parasite density.

Table 1: Summary of datasets used for training and evaluation. *Used only for external evaluation of the parasite detection.

| Dataset | Negative samples | Positive samples | Fields per sample | Annotated Parasites |
|---|---|---|---|---|
| Ibadan (Manescu et al., 2020) | 144 | 155 | 100 | 2,986 |
| Chittagong-1 (Kassim et al., 2021) | 50 | 150 | 20 | 25,765 |
| Chittagong-2* (Yang et al., 2019) | 0 | 150 | 20 | 84,961 |

### 2.2. Sample-level predictions with Multi-Instance Learning

Each TBF sample is treated as a *bag* $B = \{x_1, x_2, \ldots, x_n\}$ of image patches (instances) randomly cropped from digitized image fields acquired under a 100x 1.4 NA objective. According to MIL assumptions:

$$y_B = \begin{cases} 1 & \text{if at least one instance } x_i \text{ in bag } B \text{ contains malaria parasites} \\ 0 & \text{if all instances in bag } B \text{ are negative.} \end{cases}$$

Each tile $x_i$ is passed through a ConvNet encoder (VGG-16) (Simonyan and Zisserman, 2014) $f_\theta(\cdot)$, followed by global average pooling (GAP), producing a feature embedding

$$h_i = \mathrm{GAP}(f_\theta(x_i)), \quad h_i \in \mathbb{R}^d. \tag{1}$$

The bag-level representation is obtained via max feature pooling:

$$H(B) = \max_{i=1,\ldots,n} h_i. \tag{2}$$

A linear classifier $g_\phi(\cdot)$ then predicts the bag label probability:

$$p(y_B = 1 \mid B) = \sigma\big(g_\phi(H(B))\big), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. We trained the model with bags of 50 randomly cropped 224x224 pixel tiles from the image fields available for each sample (one tile per image field). We adopted max-pooling to aggregate field embeddings into a bag representation because, in many cases, only a small fraction of image fields in a positive sample contain parasites. Binary cross-entropy loss was used to optimize the model weights. At test time, all available image tiles from each sample were passed through the trained model, and their aggregated bag representation $H(B)$ was used to predict the sample label. The model was fully trained for 100 epochs with an SGD optimizer and a learning rate of 0.0003. MIL provides two key signals: (i) sample-level predictions and (ii) per-field confidence scores used to trigger ICAM refinement.

### 2.3. Pseudo-annotation generation with Iterative Class Activations

While MIL enables sample-level classification, it does not localize individual parasites. To generate pseudo-annotations, for each high-confidence malaria-positive image field ($p(y_B) > 0.9$), we derive Class Activation Maps (CAMs) from the MIL classifier. Given convolutional feature maps $F \in \mathbb{R}^{C\times H\times W}$ and classifier weights $w_{c,k}$ for class $c$, the CAM is computed as

$$\mathrm{CAM}_c(i,j) = \sum_{k=1}^{C} w_{c,k}\, F_k(i,j). \tag{4}$$

CAMs highlight discriminative image regions that contribute to the positive class, but often emphasize only the most prominent parasite or artifact. To address this, we introduce Iterative Class Activation Mapping (ICAM), which progressively reveals multiple faint parasite instances through a sequence of erasure and refinement steps. At each iteration $s$, the highest CAM region is erased from the image $I^s$, and a new CAM is computed if

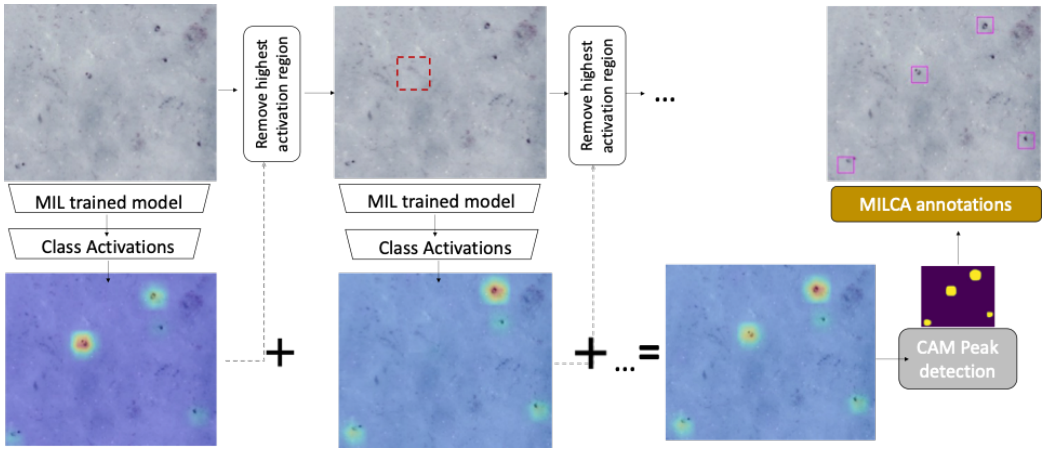



Figure 2: Iterative Class Activation Maps for Malaria Parasite Annotations Generation (ICAM).

**Algorithm 1:** Iterative Class Activation Mapping (ICAM)

**Require:** Image $I$, trained MIL classifier $f_\theta$, confidence threshold $\tau$, decay factor $\delta$, maximum iterations $S$

**Ensure:** ICAM heatmap

Initialize ICAM $\leftarrow 0$, $s \leftarrow 0$
**while** $f_\theta(I) > \tau$ **and** $s < S$ **do**
  Compute $\text{CAM}_s \leftarrow \text{CAM}(I)$
  Update ICAM $\leftarrow \text{ICAM} + \delta^{-s} \cdot \text{CAM}_s$
  $I \leftarrow \text{erase\_highest\_activation}(I, \text{CAM}_s)$
  $s \leftarrow s + 1$
**end while**
**return** ICAM

the classifier score of the modified image remains above a confidence threshold (empirically, $\tau = 0.7$). At each iteration, the ICAM map is updated by adding the new CAM with an exponential decay factor ($\delta = 2$), ensuring that earlier discoveries retain more influence:

$$\text{ICAM}(I) = \text{CAM}(I) + \sum_{s=1}^{S} \delta^{-s} \cdot \text{CAM}(I^s). \tag{5}$$

To convert ICAM maps into bounding boxes, we normalized each map, apply Otsu thresholding to isolate salient regions, eroded small artifacts with a 5x5 kernel, and detected subsequent peaks separated by at least 10 pixels. Around each peak, a bounding box of fixed size (64×64 pixels) is placed. This box dimension reflects typical parasite size in TBF microscopy (Manescu et al., 2020) . This procedure produced ∼200k pseudo-annotations on image fields from the Ibadan(Manescu et al., 2020) and Chittagong-1(Kassim et al., 2021) datasets.

**Hard negative mining.** Negative samples often contain parasite-like artifacts such as platelets and stain precipitates. We performed hard-negative mining by applying the MIL classifier to fields from negative bags and computing CAMs with the positive-class weights. Although these fields do not contain any parasites, the resulting CAMs identify texture patterns the model mistakenly finds discriminative (Sup. Fig. 1). Peaks extracted from these CAMs are converted into bounding boxes using the same procedure as ICAM. This procedure yielded ∼60k hard negative annotations.

### 2.4. Parasite detector training and pseudo-label refinement

The set of ICAM-generated parasite and artefact pseudo-annotations were next used to train an initial one-stage RetinaNet with a ResNet-50 backbone object detector (Lin et al., 2017) $\mathcal{D}_{\text{MILCA-raw}}$. We further incorporated a bootstrapped refinement stage. $\mathcal{D}_{\text{MILCA-raw}}$ is applied to all training fields, and its predictions are retained if their confidence exceeds 0.7. A second detector, $\mathcal{D}_{\text{MILCA}}$, was next trained on the refined pseudo-labels.

The MILCA detectors were trained with a focal loss (Lin et al., 2017) for 30 epochs using an Adam Optimizer with learning rate of $5 \times 10^{-4}$.

### 2.5. Fine-Tuning with Limited Manual Annotations

To evaluate annotation efficiency, we further fine-tuned $\mathcal{D}_{\text{MILCA}}$ using varying fractions of the available manually annotated data (5-75%). For each fraction, we randomly sampled the corresponding subset of annotations three times and report the mean performance across the three runs to reduce sensitivity to the particular subset chosen. Fine-tuning updates only the classification and regression heads and is performed for 10 epochs.

This hybrid regime differs from conventional semi-supervised pipelines, which require manual labels to bootstrap the initial detector. In contrast, MILCA first constructs a detector entirely from weak supervision and incorporates expert annotations only at the final refinement stage.

We compared MILCA fine-tuning with a semi-supervised learning (SemiSL) baseline based on naive pseudo-labeling previously used in microscopy (Shokrollahi et al., 2024). In this setup, a vanilla object detector was first trained on the same fractions of expert annotations as MILCA. This detector was then applied to the remaining unlabeled training images, and only predictions with a high confidence score (above 0.7) were retained as pseudo-annotations. A new detector was subsequently trained from scratch using both the original manual annotations and the generated pseudo-labels.

## 3. Results

### 3.1. Sample-level predictions

MILCA's MIL classifier accurately distinguished malaria-positive from negative samples across all three datasets. In 3-fold cross-validation, it achieved an average AUC of $0.969 \pm 0.013$ (Fig. 3a), with overall accuracy exceeding 90%. Performance was lower on low-parasitemia samples (<1000 parasites/$\mu$l), reflecting the inherent difficulty of detecting sparse parasite instances (Fig. 3b). These sample-level predictions are produced without any manual parasite annotations and provide the field-selection signal used by ICAM.

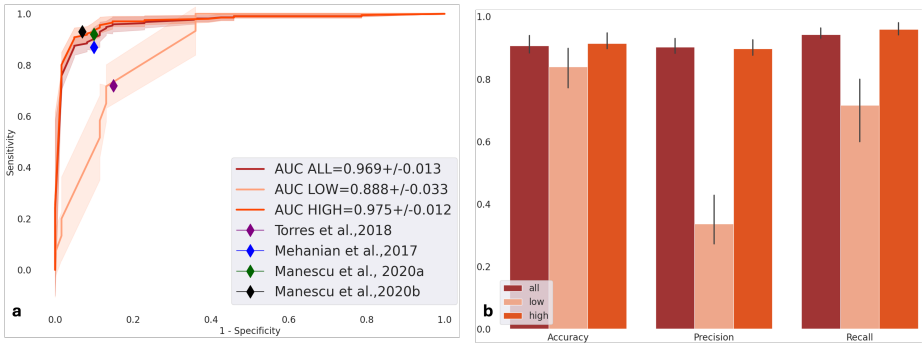



Figure 3: Sample-level classification results. (a) ROC curve with mean AUC over 3 folds ($n = 99$). LOW: <1000 parasites/$\mu$l. Comparison with previous studies is just indicative as the train or test datasets were not made available.(b) Additional sample-level metrics.

### 3.2. Weakly supervised parasite detection

Across the Ibadan dataset (23 test fields, ∼700 parasites), Chittagong-1 dataset (10,598 parasites in test set), and external Chittagong-2 dataset (84,961 parasites) a detector trained solely on ICAM pseudo-labels (no manual boxes) achieved mean AP values between 0.1–0.3 at IoU = 0.5 demonstrating that sample-level supervision alone provides sufficient signal to bootstrap a malaria parasite detector. As expected given the small parasite size, AP was higher at lower IoU thresholds (Fig. 5c). Qualitative examples illustrate that MILCA detects many parasites without supervision, and fine-tuning improves recall and bounding-box accuracy (Fig. 5a-b).

### 3.3. Label efficiency and hybrid fine-tuning

We next evaluated parasite detection using an object detector trained on MILCA pseudo-annotations and fine-tuned with different fractions of manually labeled data (5-75%). We compared against a fully supervised detector trained on all ground-truth annotations (vanilla).

Fine-tuning with small fractions of manual annotations yielded substantial gains (Fig. **??** and Fig. 4). Across Ibadan, Chittagong-1, and the external Chittagong-2 dataset, MILCA outperformed both the fully supervised detector and a naive pseudo-labeling SemiSL baseline at low budgets (For more details, see Sup. Fig. 3). With only 5% of the labels, MILCA improved AP by 10–20 points over supervised training, reflecting the stronger initialization produced by ICAM pseudo-labels. At 50% labels, MILCA remained competitive or superior across all dataset (Table 2). Evaluated on Chittagong-2—unseen during training—MILCA generalized better than supervised or SemiSL detectors. This suggests that training on large numbers of weak, diverse pseudo-labels provides robustness to staining variability and acquisition differences, which are common in routine malaria microscopy.

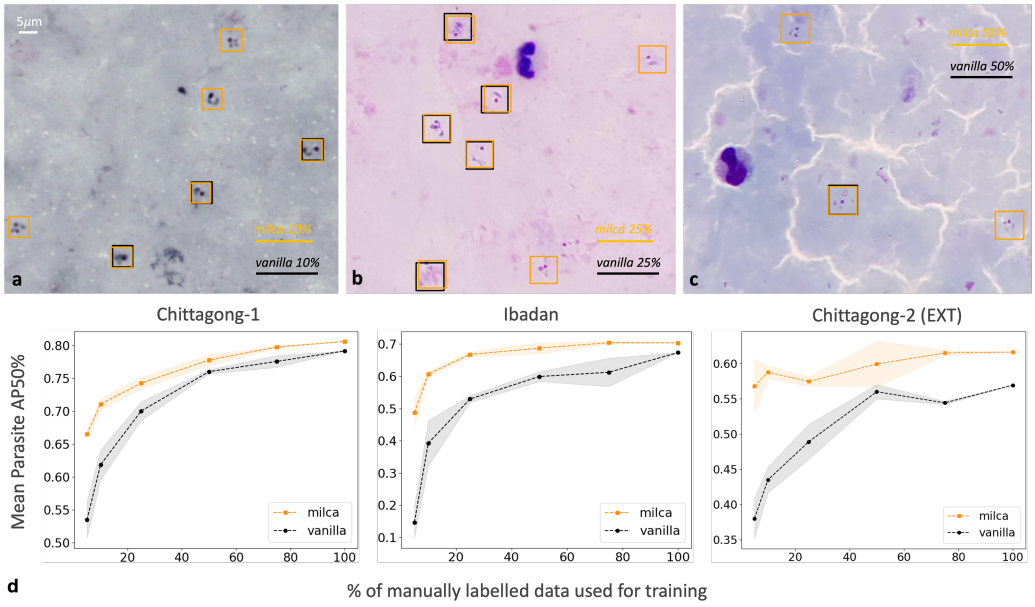



Figure 4: Example detections using the Vanilla and the MILCA and fine-tuned parasite detectors on (a) Ibadan, (b) Chittagong-1 and (c) Chittagong-2 test images. (d) Parasite detection performance. Mean AP at IoU=0.5 for different fractions of labeled data on Chittagong-1, Ibadan, and Chittagong-2 datasets. Experiments repeated 3 times with random fractions.

Table 2: Detection performance (Parasite AP50%, mean with std in parentheses) for Vanilla Supervised, SemiSL (Shokrollahi et al., 2024), and MILCA across three datasets under 5% and 50% annotation budgets. Best performance for each dataset and budget is highlighted in bold.

| Annotation Budget | Method | Ibadan (Internal) | Chittagong-1 (Internal) | Chittagong-2 (External) |
|---|---|---|---|---|
| 5% | Vanilla Supervised | 0.15 (0.05) | 0.53 (0.03) | 0.38 (0.03) |
| | SemiSL | 0.31 (0.05) | 0.56 (0.03) | 0.33 (0.02) |
| | MILCA | **0.49 (0.04)** | **0.66 (0.01)** | **0.57 (0.04)** |
| 50% | Vanilla Supervised | 0.60 (0.02) | 0.76 (0.00) | 0.56 (0.01) |
| | SemiSL | 0.64 (0.01) | 0.75 (0.01) | 0.51 (0.01) |
| | MILCA | **0.69 (0.00)** | **0.78 (0.01)** | **0.60 (0.03)** |

### 3.4. Parasite count estimation

MILCA produced accurate parasite counts on internal test sets, achieving $R^2 = 0.80$ on Chittagong-1 (Fig. 6). Performance decreased on the external Chittagong-2 dataset ($R^2 = 0.14$), consistent with known domain shift effects, yet MILCA remained more ac-

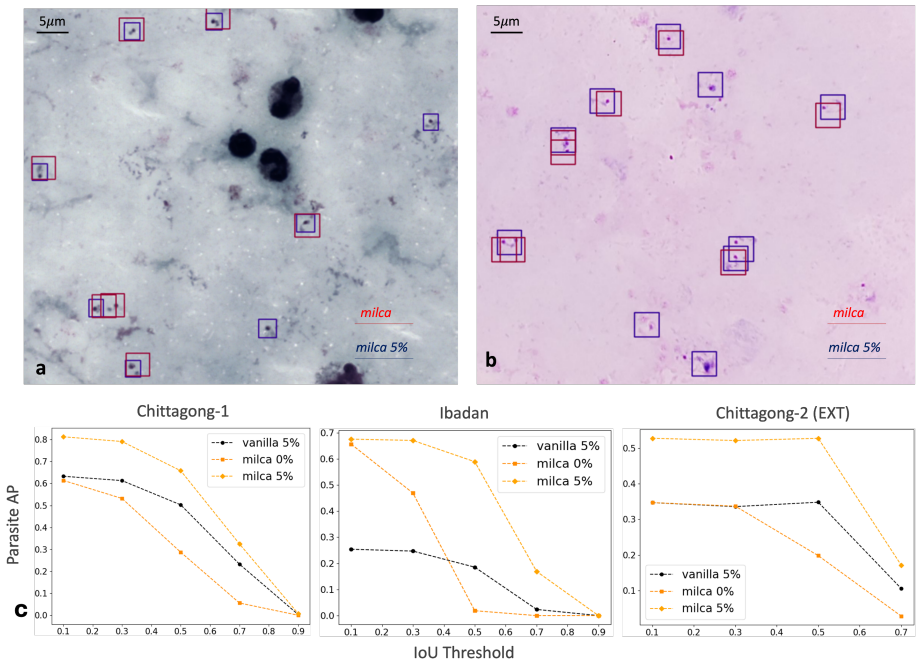



Figure 5: Example detections using the MILCA object detector (red: no fine-tuning; blue: with 5% fine-tuning) on (a) Ibadan and (b) Chittagong-1 test images.(c)Parasite detection performance. AP as a function of IoU threshold.

curate than the vanilla detector under all annotation budgets. Both models tended to underestimate counts at very high parasitemia levels, where overlapping parasites introduce ambiguity. Overall, MILCA demonstrates that meaningful object-level supervision can be derived entirely from sample-level diagnostic labels, enabling strong detection and counting performance with minimal expert annotation.

## 4. Discussion

We introduced MILCA, a weakly supervised framework for malaria parasite detection in thick blood films that relies solely on sample-level diagnostic labels. By combining MIL field selection, ICAM refinement, and hard-negative mining, MILCA generates dense pseudo-annotations that enable training a practical detector without any bounding-box labels. Across multiple datasets, MILCA achieves competitive detection and counting performance, under fully weak supervision and substantially outperforms supervised and SemiSL baselines when annotation budgets are low. A key finding is that weak supervision *before* strong supervision is highly effective: detectors initialized from ICAM pseudo-labels require far fewer manual annotations to reach high accuracy compared to conventional semi-supervised pipelines. This offers a scalable path for developing diagnostic models in settings where expert time is limited and annotation costs are prohibitive.

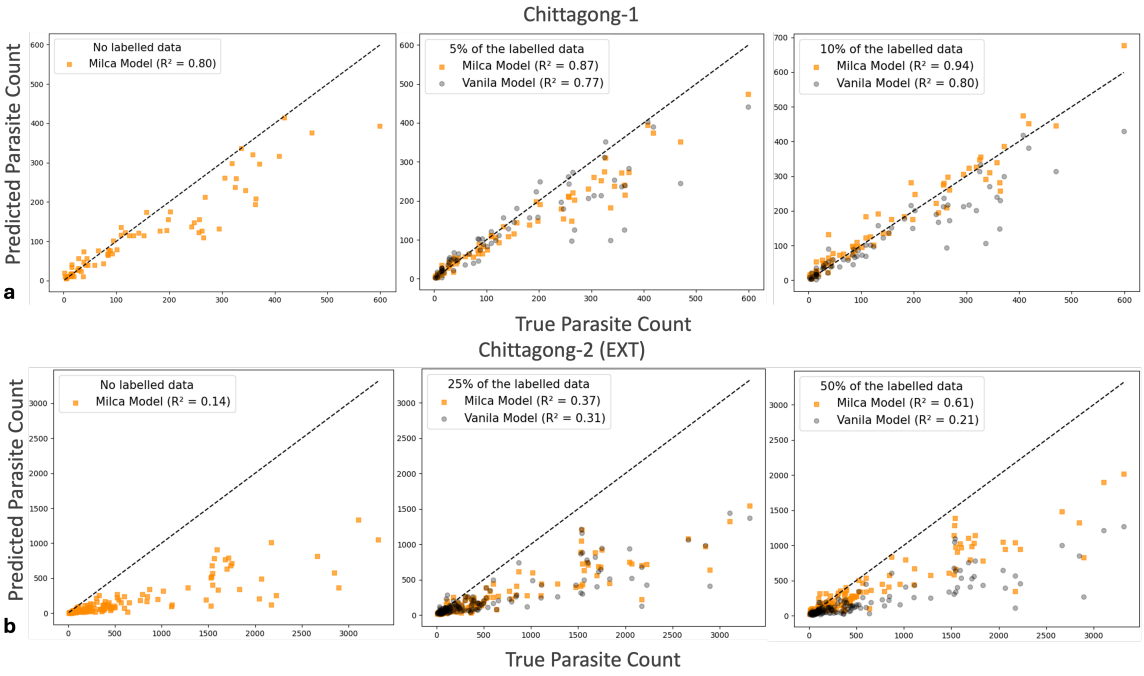



Figure 6: Parasite count evaluation on (a) Chittagong-1 (internal) and (b) Chittaging-2 (external). Scatter plots of predicted vs. true counts for MILCA with and without fine-tuning. MILCA outperforms vanilla detectors under the same annotation budget.

**Future work.** Enhancements may include incorporating uncertainty-aware pseudo-label filtering, teacher–student pseudo-labeling, joint MIL–detector training and contrastive pre-training to mitigate domain shift. MILCA could also be extended to other pathology tasks where annotations are scarce but sample-level labels are readily available.

## References

Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac mr image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 253–260. Springer, 2017.

Soufiane Belharbi, Mohammadhadi Shateri, Luke McCaffrey, Eric Granger, et al. Pixelcam: Pixel class activation mapping for histology image classification and roi localization. In *Medical Imaging with Deep Learning*, 2025.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(3):1327–1347, 2022.

Samson Chibuta and Aybar C Acar. Real-time malaria parasite screening in thick blood smears for low-resource setting. *Journal of digital imaging*, 33(3):763–775, 2020.

Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, 2024.

Zeyu Gao, Anyu Mao, Kefei Wu, Yang Li, Liebin Zhao, Xianli Zhang, Jialun Wu, Lisha Yu, Chao Xing, Tieliang Gong, et al. Childhood leukemia classification via information bottleneck enhanced hierarchical multi-instance learning. *IEEE Transactions on Medical Imaging*, 42(8):2348–2359, 2023.

Yasmin M Kassim, Feng Yang, Hang Yu, Richard J Maude, and Stefan Jaeger. Diagnosing malaria patients with plasmodium falciparum and vivax using deep learning for thick smear images. *Diagnostics*, 11(11):1994, 2021.

Hannah Kniesel, Leon Sick, Tristan Payer, Tim Bergner, Kavitha Shaga Devan, Clarissa Read, Paul Walther, and Timo Ropinski. Weakly supervised virus capsid detection with image-level annotations in electron microscopy images. *arXiv preprint arXiv:2508.00563*, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Ali Madani, Mehdi Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*, pages 1038–1042. IEEE, 2018.

Petru Manescu, Michael J Shaw, Muna Elmi, Lydia Neary-Zajiczek, Remy Claveau, Vijay Pawar, Iasonas Kokkinos, Gbeminiyi Oyinloye, Christopher Bendkowski, Olajide A Oladejo, et al. Expert-level automated malaria diagnosis on routine blood films with deep neural networks. *American Journal of Hematology*, 95(8):883–891, 2020.

Petru Manescu, Priya Narayanan, Christopher Bendkowski, Muna Elmi, Remy Claveau, Vijay Pawar, Biobele J Brown, Mike Shaw, Anupama Rao, and Delmiro Fernandez-Reyes. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. *Scientific Reports*, 13(1):2562, 2023.

Courosh Mehanian, Mayoore Jaiswal, Charles Delahunt, Clay Thompson, Matt Horning, Liming Hu, Travis Ostbye, Shawn McGuire, Martha Mehanian, Cary Champlin, et al. Computer-automated malaria diagnosis and quantitation using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 116–125, 2017.

World Health Organization and Center for Disease Control. *Basic malaria microscopy*. World Health Organization, 2010.

World Health Organization et al. *World malaria report 2022*. World Health Organization, 2022.

Yasin Shokrollahi, Karina Pinao Gonzales, Maria Esther Salvatierra, Simon P Castillo, Tanishq Gautam, Pingjun Chen, B Leticia Rodriguez, Sara Ranjbar, Patient Mosaic Team, Luisa Solis Soto, et al. Advancing multiplex immunofluorescence imaging cell detection using semi-supervised learning with pseudo-labeling. In *Medical Imaging with Deep Learning*, pages 1448–1461. PMLR, 2024.

Lambok Siahaan. Laboratory diagnostics of malaria. In *IOP Conference Series: Earth and Environmental Science*, volume 125, page 012090. IOP Publishing, 2018.

John-William Sidhom, Ingharan J Siddarthan, Bo-Shiun Lai, Adam Luo, Bryan C Hambley, Jennifer Bynum, Amy S Duffield, Michael B Streiff, Alison R Moliterno, Philip Imus, et al. Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. *NPJ precision oncology*, 5(1):38, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Katherine Torres, Christine M Bachman, Charles B Delahunt, Jhonatan Alarcon Baldeon, Freddy Alava, Dionicia Gamboa Vilela, Stephane Proux, Courosh Mehanian, Shawn K McGuire, Clay M Thompson, et al. Automated microscopy for routine malaria diagnosis: a field comparison on giemsa-stained blood films in peru. *Malaria journal*, 17(1):339, 2018.

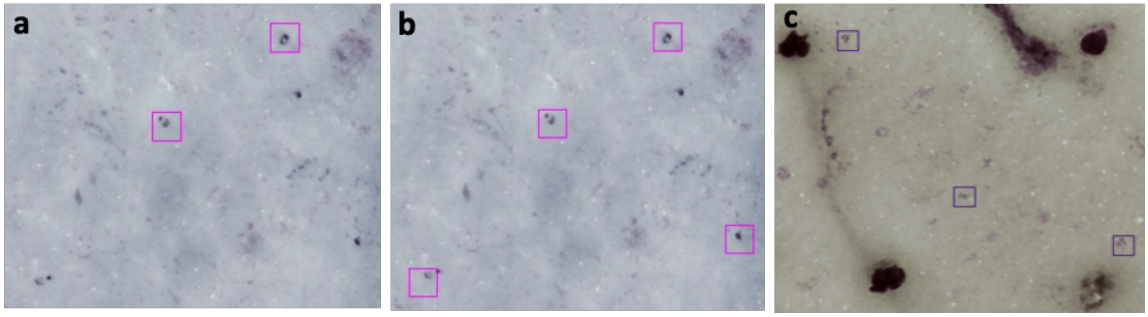
Mou-Cheng Xu, Yukun Zhou, Chen Jin, Marius de Groot, Daniel C Alexander, Neil P Oxtoby, Yipeng Hu, and Joseph Jacob. Bayesian pseudo labels: Expectation maximization for robust and efficient semi-supervised segmentation. In *International Conference on*

*Medical Image Computing and Computer-Assisted Intervention*, pages 580–590. Springer, 2022.

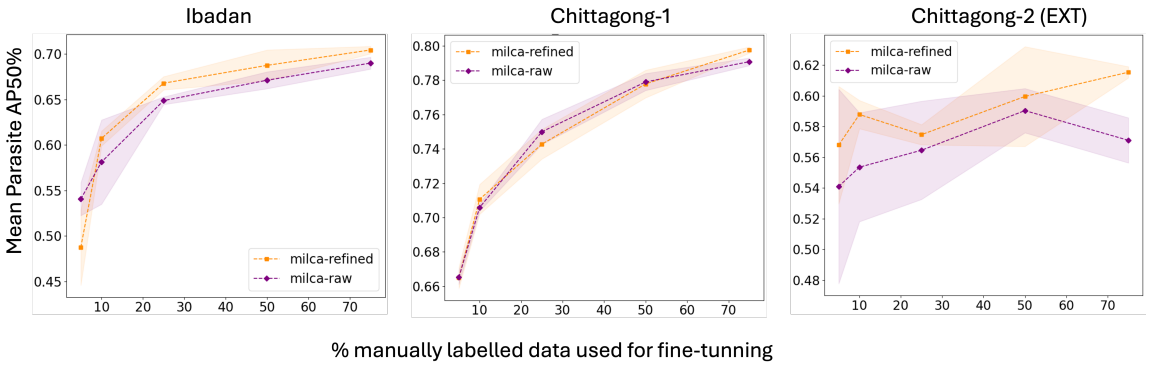
Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard J Maude, Stefan Jaeger, and Sameer Antani. Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE journal of biomedical and health informatics*, 24(5):1427–1438, 2019.

Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5866–5885, 2021.

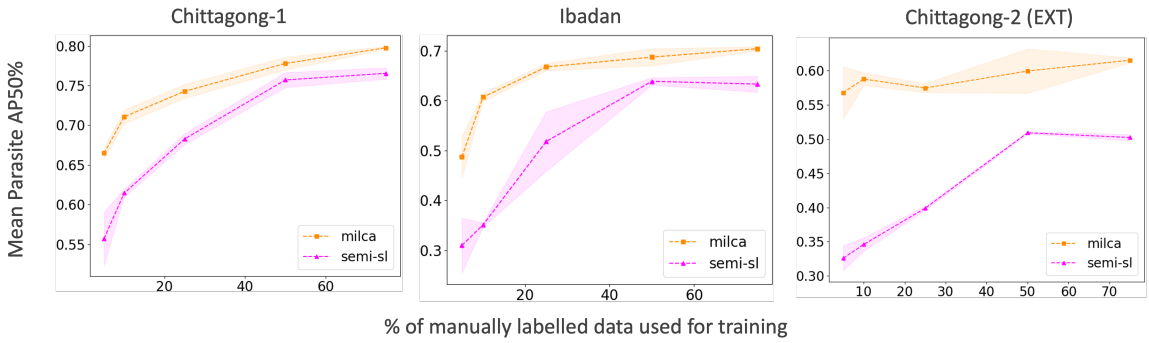
## Appendix A. Supplementary Figures



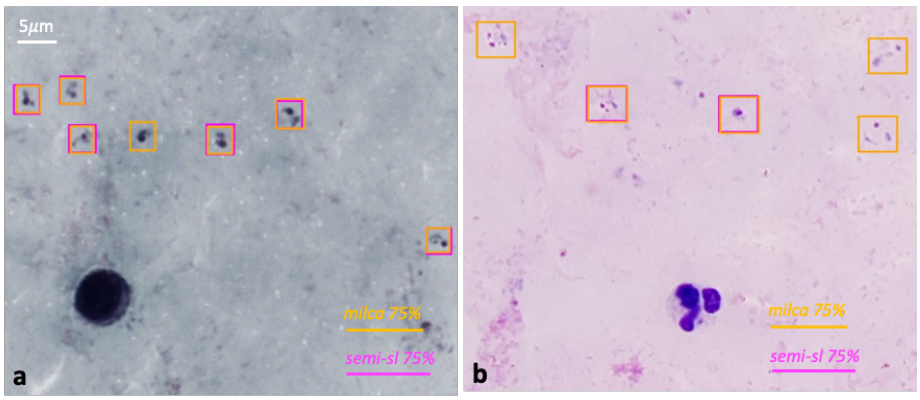

Sup. Fig. 1: (a) Malaria parasite partial annotations generated with CAM. (b) Malaria annotations augmented with ICAM. (c) Hard negative annotations generated by CAM.




Sup. Fig. 2: Parasite detection performance. Mean AP at IoU=0.5 for different fractions of labeled data on Chittagong-1, Ibadan, andChittagong-2 datasets. Experiments repeated 3 times with random fractions. Milca-refined refers to the bootstrapped pseudo-label retrained model.




Sup. Fig. 3: Comparison with a naive pseudo-labeling SSL baseline. MILCA fine-tuning achieves higher AP across all fractions of labeled data.




Sup. Fig. 4: Example detections using MILCA fine-tuned and the Semi Supervised object detectors on (a) Ibadan and (b) Chittagong-1 test images.